ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Geometric Data Transformation for Privacy Preserving On Data Stream Using Classification

Krupali N. Vachhani, Dinesh B. Vaghela

Student, Dept of C.S.E., PIT, Gujarat Technological University, Gujarat, India

Assistant Professor, Dept. of C.S.E., PIT, Gujatat Technological University, Gujarat, India

ABSTRACT: Data mining is an information technology that extracts valuable knowledge from large amounts of data. Recently, data streams are emerging as a new type of data, which are different from traditional static data. Data flows in and out with fast speed; and immediate response is required. Traditional algorithm is designed for the static database. If the data changes, it would be necessary to rescan the database, which leads to long computation time and inability to promptly respond to the user. To preserve data privacy during data mining, the issue of privacy preserving data mining has been widely studied and many techniques have been proposed. However, existing techniques for privacy-preserving data mining are designed for traditional static databases and are not suitable for data streams. So the privacy preservation issue of data streams mining is a very important issue. This work is about proposing a Method and algorithms for the process of Geometric Data Perturbation or Geometric Data Transformation to achieve privacy preservation. Geometric Data Transformation is a kind of data perturbation techniques. In this, we describe the geometric transformations including translation, scaling, rotation, which can transform data in the protection of privacy while maintaining the similarity between data objects. So, it maintain data utility and accuracy. Our main goal is to preserve the privacy along with accuracy.

KEYWORDS- DATA: Stream Mining, Geometric Data Transformation, Privacy Preserving.

I. INTRODUCTION

Data mining is nothing but extracting meaningful knowledge from the large amount of data. We can classify data mining techniques as follows: classification, association rule mining, clustering, sequential pattern analysis, data visualization, prediction. In recent years, simple transactions like using credit card, browsing the web, phone database, sensor network lead to wide and automated data storage. All these have large flows of data continuously and dynamically. This type of large volume data leads to many mining and computational challenges. Our main goal is to preserve the privacy along with accuracy. Accuracy should be maintain because if information or data loss is more, then no meaning of privacy. So, we will use some privacy preserving techniques to transform original data to transformed data.

A. DATA STREAM MINING

Data stream is new type of data that is different than traditional static database. Data stream is continuous and dynamic flow of data. Data streams are emerging as a new type of data, which are different from traditional static data. The characteristics of data streams are: Data has timing preference; data distribution changes constantly with time; the amount of data is enormous; Data flows in and out with fast speed; and immediate response is required. Traditional algorithm is designed for the static database. If the data changes, it would be necessary to rescan the database, which leads to more computation time and inability to promptly respond to the It is sequence of real time data with high data rate and application can read once.

We have many data mining algorithm for traditional database where data is static and continuous flow. Use of traditional data mining algorithm is not appropriate in data stream mining because of no control over dataflow. If data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

will change, then we have to rescan the database. This will take more computational time. In data stream mining data is not persistent but rapid and time varying. Once element of data stream is processed, it is discarded. So, it is not easy to retrieve it unless if we explicitly store them in memory.

B. NEED FOR PRIVACY IN DATA MINING

Generally privacy means "keep information about me from being available to others". The main goal is information not be misused. Because if once information is released, it will be impossible to prevent misuse. There is no problem if someone knowing my birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft. Advantages of Privacy Protection are protection of personal information, protection of proprietary or sensitive information, enables collaboration between different data owners without reveal their information to each other. Informational privacy is related to the manner in which personal information is collected, used and disclosed. Some issues are there like data quality, accuracy, utility.

II. RELATED WORK

Agrawal and Srikant [Agrawal and Srikant 2000] considered the case of building a decision-tree classifier from training data in which the values of individual records have been perturbed, by adding random values from a probability distribution. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. While it is not possible to accurately estimate original values in individual data records, they proposed a novel reconstruction procedure to accurately estimate the distribution of original data values. The distribution reconstruction process naturally leads to some loss of information. In [Agrawal and Agrawal 2001], proposed a new algorithm for distribution reconstruction which is more effective than that proposed in [Agrawal and Srikant 2000], in terms of the level of information loss. This algorithm, based on Expectation Maximization (EM) algorithm, converges to the maximum likelihood estimate of the original distribution based on the transformed data, even when a large amount of data is available. They also pointed out that the EM algorithm was in fact identical to the Bayesian reconstruction proposed in [Agrawal and Srikant 2000], except for the approximation partitioning values into intervals. According to the way training data are obtained, the construction of a classification model can be distinguished into non-incremental learning and incremental learning. Domingos and Hulten proposed the VFDT(Very Fast Decision Tree Learner) algorithm to solve the problem of long learning time. The VFDT algorithm belongs to the third category of incremental learning and uses the statistical results of the Hoeffding bounds to determine using fewer samples if the difference between the gain value of the best attribute and that of the second best test attribute is greater than a deviation value. When it is the case, it indicates that the best test attribute in the sample data can be used as the best test attribute of the whole data. Using this attribute as the test attribute in the root node, the remaining data are mapped to the leaf nodes according to the test in the root node and are used to select the test attributes in the leaf nodes. The main drawback of the VFDT algorithm is its inability to handle data distribution from different time. For many applications, new data are usually more important than old data. The VFDT algorithm does not consider the time of data, and hence cannot mine data from different time.

III. CLASSIFICATION ON DATA STREAM

Many government organizations, businesses and non-profit agency to Caring their short-and long-term schedule activities, to gather for a way to store, analyze and report data on persons, households or businesses looking. Confidential information such as social numbers, income, credit ratings, type of illness, customer purchases, etc., that needs to be sufficiently endangered. Classification is the process of resulting a model that describe and decides data classes or concepts, for the purpose of being able in the direction of use the model to predict the class of objects whose class label is not known. This approach focused on the overall quality of generated classifier after dataset transformation.

Privacy-preserving classification of data streams Method (PCDS)[8] is the process of data streams classification to get enough privacy conservancy. This method is classified into two stapes, 1) **Data streams pre-processing** and 2) **Data streams mining**.

Data streams pre-processing: For data excruciating and transformation algorithm to perturb personal data. Users can flexibly adjust the data attributes. So threats and risks from releasing data can be effectively reduced.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Data streams mining: The weighted average sliding window (WASD) algorithm to mine transformed data streams. When the classification error proposition exceeds a predetermined value, the classification model is reconstructed to maintain classification accuracy.

IV. PROPOSED ALGORITHM

A. *Description of the Proposed Algorithm:*

Aim of the proposed algorithm is to get approximate equal accuracy of modified dataset as per original dataset. The algorithm is given below.

For each attribute of G(X), let R be random rotation ,X be a original dataset, T be a translation and D be a Gaussian noise then the value of attribute G(X) is calculated using following formula.

$$G(X) = R*X + T + \Delta$$

Procedure: Geometric Transformation Based Multiplicative Data Perturbation.
Input: Data Stream D, Sensitive attribute S.
Intermediate Result: Transformed data stream D'.
Output: Clustering results R and R' of Data stream D and D' respectively.
Steps:

1. Given input data D ith tuple size n, extract sensitive attribute $[S]_{n\times 1}$.
2. Rotate $[S]_{n\times 1}$ into 180o clock-wise direction and generate$[R_S]_{n\times 1}$.
3. Multiply elements of [S] with $[R_S]$, transformed sensitive attribute values will be $[X]_{n\times 1}= [S]_{n\times 1} \times [R_S]_{n\times 1}$
4. Calculate translation T as mean of sensitive attribute $[S]_{n\times 1}$.
5. Generate transformation $[St]_{n\times 1}$ by applying translation T to $[S]_{n\times 1}$.
6. Calculate Gaussian distribution P(x) as a probability density function for Gaussian noise $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ where, μ=Mean, σ=Variance
7. Geometric data perturbation of sensitive attribute $[Gs]_{n\times 1} = [X]_{n\times 1} + [St]_{nx1} + P(x)$.
8. Crate transformed dataset D' by replacing sensitive attribute $[S]_{n\times 1}$ in original dataset D with $[Gs]_{n\times 1}$.
9. Apply classification algorithm with different values of k on original dataset D having sensitive attribute S.
10. Apply classification algorithm with different values of k on perturbed dataset D' having transformed sensitive attribute Gs.
11. Create cluster membership matrix of results from step 9 and step 10 and analyze.

Where, $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$: Probability Density Function.

V. PROPOSED WORK

Here the idea is to extend the traditional data mining technique to stream data for sensitive information. Main problem is how to modify the data and how to recover data from the result. We are doing transformation of given dataset D and will get new modified dataset D' .Here we compare original dataset and modified data set for accuracy that they must have approximate equal. We compare them in terms of less information loss, response time, and more privacy gain so get better accuracy. We have used MOA tool rather than weka tool. Weka tool works on static dataset where data remain unchanged but Moa tool works on dynamic dataset where data is continuous changes. We performed transformation using geometric data transformation. Our main goal is to maintain the accuracy along with privacy and we are focusing on trying to get approximately same accuracy as original data has.

A. ROTATION TRANSFORMATION

Rotation transformation is used for classification and clustering. for privacy preservation. Rotation transformation is our initial step. Rotation transformation is shown as G(x) =RX, where R is rotational matrix. We can perform rotation by providing angle in two dimension or by clock wise or anti clockwise direction. Here we are using clock

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

wise 180˚ rotation. Input should be .csv file and the multiple columns which we want to rotate. Output of rotation should be multiple column with changed value.

B. TRANSLATION

In translation we are translating the original value based on mean. For translation first we have to find out the mean of original dataset. Then performing translation by addition of mean with original value and the result is translated value as output.

C. GEOMETRIC DATA TRANSFORMATION

Geometric data transformation is very popular data transformation technique. It is a multiplicative data transformation which is used in privacy preserving in collaborative data mining. If we compare this Geometric data transformation to other data transformation techniques, geometric data transformation has many advantages over privacy preservation so it is most widely used. Many popular data mining models are invariant in geometric transformation. Another advantage is geometric data transformation over other method is its low cost and maintain the accuracy of sensitive data. As compared to other approaches, geometric data transformation reduces the complexity in balancing data utility and data privacy guarantee [5]. While preserving the privacy guarantee some technique lose the data utility of sensitive information. Geometric transformation transformed the dataset before releasing to use for public. Geometric data transformation is combination of rotation transformation and translation transformation and noise addition.

$$G(X)=RX+T+\Delta$$

Gaussian noise $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

where, μ=Mean, σ=Variance, x=original data

R is rotation transformation, X is original dataset, T is translation transformation and Δ is Gaussian noise addition. Privacy is maintained by noise addition which gives protection to data and stop data being disclosed. Addition of noise is possible to categorical data and also numerical data. Using random noise is not appropriate because we may lose the data or value of sensitive attributes then we have to calculate that value as null or average so that affect the result and also give less accuracy.

VI. SIMULATION RESULTS

This technique is developed in java. After implementation we are using Moa tool for providing stream environment and to measure the parameters for performance. We have taken two real dataset from UCI machine learning repository. Detailed about dataset is given in below table. Bank management dataset contain total 45,211 instances and we are applying transformation on age and duration. Adult dataset contains total 32,561 instances and we are applying transformation on age and education number. After applying noise to the data we will check the performance measure for both original and transformed data. For analysis we applied two different classification algorithms that is naive Bayesian and J48 algorithm. By applying both algorithms on each attribute of adult dataset we can find out the results shown in table 2, 3, 4 and table 3. From the result table we can also graphical representation for each attribute and we can compare that which algorithm gives better result. For graph we have taken correctly classified instances on x and applied algorithm on y axis. For each table, graph is generated for evaluation.

The generated result is for the Adult dataset is shown below in table. In adult dataset we have processed two attributes Age and Education Num. Both original and transformed dataset attributes are passed from Classification algorithms J48 and NavieBayes (NB) and generate the efficient result. It shows the fairly good level of result that can help us while mining of data for preserving privacy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Table 1: Result of Adult dataset

	Adult - Age				Adult - Education Num			
	NB		J48		NB		J48	
	Original	Perturb	Original	Perturb	original	Perturb	Original	Perturb
correctly classified instances	0.8342	0.8318	0.8621	0.8573	0.8342	0.8291	0.8621	0.8562
incorrectly classified instances	0.1657	0.1681	0.1378	0.1426	0.1657	0.1708	0.1378	0.1437
Time taken	0.2	0.23	4.52	4.18	0.2	0.22	4.84	5.04
Kappa statistic	0.4993	0.4905	0.6004	0.5805	0.4993	0.475	0.6004	0.5721
Mean Absolute Error	0.1735	0.1759	0.1942	0.2009	0.1735	0.1771	0.1942	0.2031
Root Mean Squared Error	0.3723	0.3756	0.3196	0.3246	0.3723	0.3152	0.3196	0.3297
Relative Absolute Error	0.4745	0.4809	0.5309	0.5495	0.4745	0.4844	0.5309	0.5553
Root Relative Squared Error	0.8706	0.8783	0.7474	0.7592	0.8706	0.8772	0.7474	0.7711

Table 2 : Result of Bank dataset

	Bank - age				Bank - Duration			
	NB		J48		NB		J48	
	original	Perturb	Original	Perturb	original	Perturb	original	Perturb
correctly classified instances	0.8807	0.8805	0.9031	0.903	0.88	0.866	0.9031	0.8924
incorrectly classified instances	0.1193	0.1195	0.0968	0.0969	0.1193	0.1339	0.0968	0.1075
Time taken	0.43	0.45	6.5	7.17	0.44	0.46	7.72	7.94
Kappa statistic	0.4391	0.4346	0.4839	0.4846	0.4391	0.3413	0.4839	0.3354
Mean Absolute Error	0.1532	0.1542	0.1269	0.1276	0.1532	0.1681	0.1269	0.157
Root Mean Squared Error	0.3088	0.3075	0.2773	0.2781	0.3088	0.3305	0.2773	0.2986
Relative Absolute Error	0.7416	0.7464	0.6142	0.6175	0.7416	0.8135	0.6142	0.7596
Root Relative Squared Error	0.9606	0.9567	0.8628	0.8653	0.9606	1.028	0.8628	0.9289

Below shows the graph of comparison with original and transformed dataset when we apply both J48 and NB algorithm to both original and transformed data. In all figure red column shows transformed data and blue column shows original data. On Y axis I have taken algorithm and on X axis I have taken accuracy. Figure 1 shows the result of correctly classified instance of adult age. Figure 2 shows correctly classified instance of adult education number. Figure 3 shows result of correctly classified instance of age from bank dataset. Figure 4 shows correctly classified instance of duration from bank dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

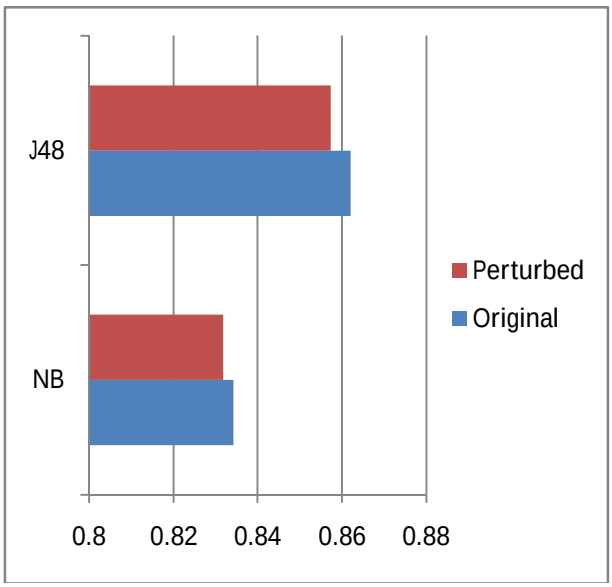

Fig.1- Correctly Classified instances Adult (Age)

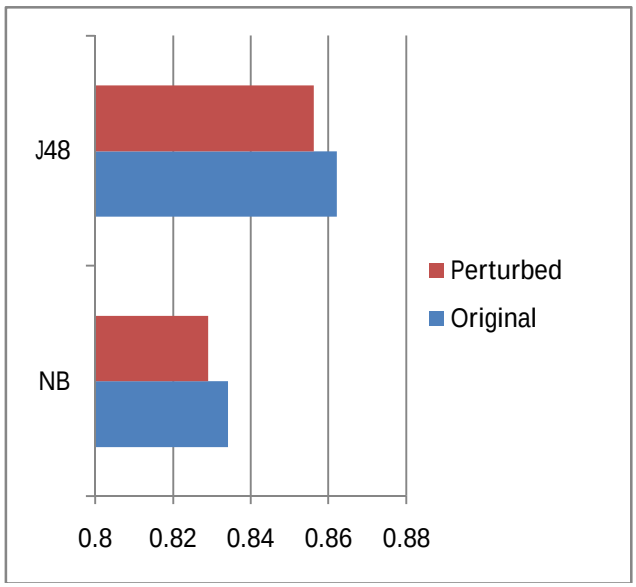

Fig. 2. Correctly Classified instances Adult (Education Num)

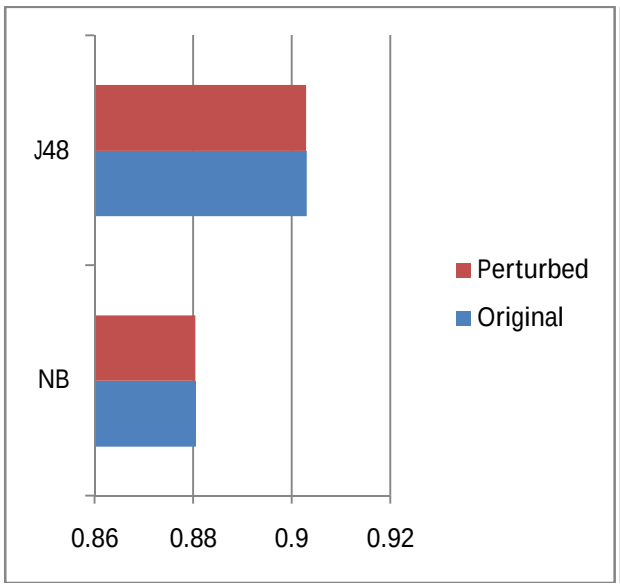

Fig. 3. Correctly Classified instances Bank (Age)

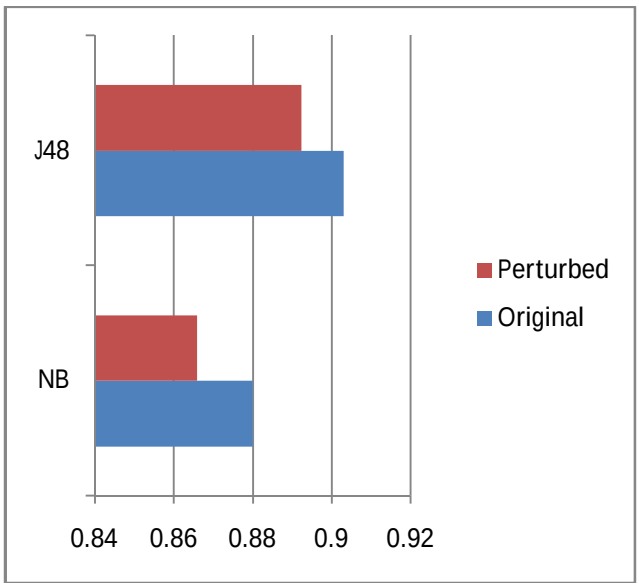

Fig 4. Correctly Classified instances Bank (Duration)

VII. CONCLUSION AND FUTURE WORK

Our requirement is more privacy and without more data loses and also maintains accuracy. So, to overcome with this problem I have used geometric transformation technique with noise addition, Gaussian noise to perturb data which provide less losses in data. Privacy must be increase. Maintain the accuracy. With Geometric Data Perturbation technique we worked on numeric data only. So we can add this algorithm on non -numeric dataset using k-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

anonymization techniques. Also we can expand algorithm to check real time data using Stream analysis tool. Also this algorithm is applied to more classification as well as clustering algorithm.

REFERENCES

1. Golab, L. And Ozsu, M.,"Issues in Data Stream Management," ACM SIGMOD Record, Vol. 32, pp. 5-14(2003).
2. Keke chen, Ling Liu, Privacy Preserving Multiparty Collaborative Mining With Geometric Data Perturbation, IEEE TRANSACTION ON PARALLEL AND DISTRIBUTED COMPUTING, VOL.XX, NO. XX. JANUARY 2009.
3. H. Chhinkaniwala and S. Garg," Tuple Value Based Multiplicative Data Perturbation Approach to preserve privacy in data stream mining", IJDKP,Vol3, No.3 May 2013.
4. Xinjun Qi , Mingkui Zong , "An Overview of Privacy Preserving Data Mining", ScienceDirect , 2011.
5. R. Agrawal and R. Shrikant, "Privacy preserving data mining" in Proceeding of ACM SIGMOD Conference, 2000.
6. A.Evfimievski, R. Shrikant, and J. Gehrke, "Limitimg Privacy Breaches in privacy preserving data mining", in Proceedings of ACM Conference on Principles of Database System(PODS), 2003.
7. Md Zahidul Islam , Ljiljana Brankovic," Privacy Preserving Data Mining: A Noise Addition Framework Using A Novel Clustering Technique",ScienceDirect , 2011.
8. Ching-Ming Chao, Po-Zung Chen, "Privacy-Preserving Classification of Data Streams", Tamkang Journal of Science and Engineering ,2009.
9. Ching-Ming, Po-Zung & Chu-Hao," Privacy Preserving Clustering Of Data Streams", Tamkang Journal Of Sc. & Engg,Vol.13 No. 3 Pp.349-358.
10. Keyur Dodiya, Shruti Yagnik," Classification Techniques for Geometric Data Perturbation in Multiplicative Data Perturbation",IJEDR, 2014.
11. Domingos, P. and Hulten, G., "Mining High-Speed Data Streams," Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining, pp. 71□80 (2000).
12. Stanley R. M. Oliveira1, Osmar R. Za _ane, "Privacy Preserving Clustering by Data Transformation Journal of Information and Data Management" , Vol. 1, No. 1, February 2010, Pages 37{51.